PAPER

Binaural Sound Source Localization in Noisy Reverberant Environments Based on Equalization-Cancellation Theory

Thanh-Duc CHAU[†a)], **Junfeng LI**[††b)], *Nonmembers*, *and* **Masato AKAGI**[†c)], *Member*

SUMMARY Sound source localization (SSL), with a binaural input in practical environments, is a challenging task due to the effects of noise and reverberation. In psychoacoustic research field, one of the theories to explain the mechanism of human perception in such environments is the well-known equalization-cancellation (EC) model. Motivated by the EC theory, this paper investigates a binaural SSL method by integrating EC procedures into a beamforming technique. The principle idea is that the EC procedures are first utilized to eliminate the sound signal component at each candidate direction respectively; direction of sound source is then determined as the direction at which the residual energy is minimal. The EC procedures applied in the proposed method differ from those in traditional EC models, in which the interference signals in rooms are accounted in E and C operations based on limited prior known information. Experimental results demonstrate that our proposed method outperforms the traditional SSL algorithms in the presence of noise and reverberation simultaneously.

key words: *binaural sound localization, equalization-cancellation model, noisy reverberant environments, humanoid robot*

1. Introduction

Binaural sound source localization (SSL) is the task of determining location of a sound source from a binaural input, of which one of the important applications is humanoid robot. In a human-robot communication system, source location information, or at least *direction of arrival* (DOA) of the observed sounds, is required to enable the robot to imitate some basic human behaviors, such as facing the user when it is called. SSL in such binaural systems is a challenging problem because sound signals observed at the receivers (ears or microphones) are corrupted by noise and reverberation in enclosed spaces, while the input is limited to only two channels with the effects of head, torso and outer ear, which is normally referred to as *head-related transfer functions* (HRTFs). Specifically, reverberation smears the direct-path sound in two main ways: self-masking caused by early reflections and overlap-masking caused by late reflections, in which the overlap-masking is a serious effect when reverberation is high. Together with reverberation, background noise in rooms makes the sound source more difficult to be correctly detected, especially when the noise is directional. In addition, HRTFs are dependent on the shape, size and material of the robot. This may reduce the effectiveness of general state-of-the-art SSL methods when applied in binaural systems. Although binaural SSL has been researched for many years, the problem of binaural SSL in noisy reverberant environments has still not been completely solved. The work presented in this paper aims at a binaural DOA estimation method, which is expected to work effectively on binaural systems and be robust against background noise and reverberation in rooms.

In the past decades, a large number of DOA estimation methods have been introduced [1], [2], in which each one differs from others by how localization cues are exploited. Two most important cues for localization in horizontal plane are interaural time difference (ITD) and interaural level difference (ILD). In practical conditions, these cues are corrupted by noise and reverberation. The well-known GCC-PHAT method [3], which is the combination of *generalized cross-correlation* and *phase transform* weighting, does not account well for noise. Therefore, although this method was reported with relatively good DOA estimation on reverberant signals [4], its performance is degraded when both noise and reverberation are simultaneously present. Moreover, since this method is based on ITD only, there has been analysis showing that it suffers from binaural setups [5].

In order to effectively localize sound sources with a binaural system, various azimuth-dependent models of binaural cues have been investigated. Andersson et al. [6] and Raspaud et al. [7] explicitly combined estimations of ITD and ILD. They showed that the combination of both ITD and ILD provides better azimuth estimation compared to ITD alone. Other research considered employing these cues implicitly. For example, Berglund et al. [8] extracted binaural cues in a feature vector and mapped it to source location using *artificial neural network*. In all of these methods, since the effect of interference signals has not been taken into account, their applicability in adverse noisy reverberant environments is still limited. More recently, in the work of Woodruff et al. [9], noise information was integrated into a joint ITD-ILD statistical model. They achieved significant improvement in comparison with the previous methods in experimental conditions. However, this method requires prior information of *direct-to-residual ratio* (DRR), which is normally not available in practice.

Concerning SSL with the effect of HRTF, there have been methods exploiting HRTF information directly. Keyrouz et al. [10] used the inverse of HRTF at each ear as a fil-

Manuscript received January 24, 2014.
Manuscript revised May 26, 2014.

[†]The authors are with Japan Advanced Institute of Science and Technology (JAIST), Nomi-shi, 923-1292 Japan.
[††]The author is with Institute of Acoustics, Chinese Academy of Sciences, China.
a) E-mail: duc.chau@jaist.ac.jp
b) E-mail: lijunfeng@hccl.ioa.ac.cn
c) E-mail: akagi@jaist.ac.jp

DOI: 10.1587/transfun.E97.A.2011

ter to recover the original sound emitted at the source. In this way, the pair of inverse HRTFs corresponding to the direction of sound source should provide the most identical 'recovered' signals. A similar mechanism was applied by McDonald et al. [11] in which the observed signal at an ear is filtered by the HRTF measured at the other ear. These methods, however, are highly dependent on HRTFs and suffer from reverberation since HRTFs vary largely along reverberation levels due to the presence of multiple reflections.

In psychoacoustic research field, directional hearing has been studied for more than a century and its mechanism was simulated by a number of binaural models. Two seminal binaural interaction models are the *coincidences* model of Jeffress [12] and the *equalization-cancellation* (EC) model of Durlarch [13], [14]. The others are supposed to be derived from one (or in some cases both) of these models (see [15]). The coincidences model is commonly realized as *cross-correlation* (CC) model, which was the principle of the standard GCC-PHAT [3] and a large number of GCC-based methods (e.g. [16]–[19]). The EC model was originally proposed to explain the mechanism of binaural detection in noise [13]. However, its concept can be extended to selective hearing in the presence of multiple sound signals, which is usually referred to as the 'cocktail party effect' [20]. This suggested that the EC model has potential for sound localization and segregation in the presence of multiple interference signals.

Inspired by the EC model, in this study, we propose a binaural DOA estimation method by integrating the EC procedures into a beamforming technique. Specifically, a null is steered to each candidate direction by first compensating for interaural phase difference (IPD) and ILD so as to equalize the signal components observed from that direction at both ears, then the total 'equalized' signal at one ear is subtracted from that at the other ear. Direction of sound source is determined as the candidate at which the residual energy of the null reaches the minimum. This work was partly presented in [21], namely EC-BEAM, in which we briefly introduced the idea and examined it in low noise conditions. In this paper, we refine the method with insight analysis, identify its problem in the presence of noise and reverberation, and suggest two strategies to overcome the problem. Specifically, one strategy is to deal with background noise by learning its effect on EC operations, provided that noise is stable in time and noise-only periods can be prior obtained. We account for not only diffuse noise but also directional noise, which is rarely considered in previous SSL research. The other strategy is to reduce the effect of reverberation, particularly the effect of late refection component, based on a hypothesis that late reflections are not correlated with target signal and together uncorrelated at both channels. We show that the EC-BEAM algorithm with combination of both strategies, named as Robust EC-BEAM, can perform effectively in noisy reverberant conditions.

Our proposed method shares the common point with the methods in [6]–[11] by a training step to learn the interaural differences under the effect of HRTFs. However, the proposed method is more applicable than [6]–[8], [10], [11] by taking the effect of noise and reverberation into consideration. Our algorithm is more flexible than [10], [11] as it is not strictly dependent on HRTFs. The mechanism to account for interference signals of our method is more reasonable than that of the method in [9] since DRR is not required. In addition, the proposed method is motivated by psychoacoustic model [13], [14] and its mechanism to adapt to interference effect is supported by psychoacoustic research [22], which revealed that perception of the human hearing system is improved when being in room for a relatively short time.

The remainder of this paper is organized as follows. Section 2 briefly introduces the concept of the EC model. In Sect. 3, first the basic idea of integrating the EC model into SSL (EC-BEAM) is presented; then the effect of noise and reverberation on EC-BEAM is analyzed and two strategies are suggested to overcome this problem (Robust EC-BEAM). In Sect. 4, we evaluate performance of each suggested strategy individually and further compare performance of the proposed Robust EC-BEAM algorithm with those of the well-known SRP-PHAT method [2] and the HRTF-based algorithm introduced in [11]. Discussion is given in Sect. 5 to provide an insight understanding of the proposed method, followed by a summary in Sect. 6.

2. Equalization-Cancellation Model

The equalization-cancellation (EC) model was originally developed by Durlach [13], [14]. In the original EC model, when a subject is presented with a binaural stimulus (target) masked by another one (masker), the auditory system attempts to eliminate the masking components by transforming the stimuli presented to the two ears, so as to equalize the masking components (the E operation), then subtracting (the C operation). This mechanism was originally applied to signal detection [13]. It was shown that the EC model is able to predict a large set of binaural masking level differences (BMLDs), where the BMLD is defined as the difference in the detection threshold between binaural and monaural conditions. The model was further improved by Culling and Summerfield [23], in which the EC procedures are performed independently in each frequency band. Due to its ability to explain the perception mechanism in 'cocktail party' scenarios, the EC model has been extensively utilized in a number of signal processing tasks, such as speech intelligibility prediction [24], [25], sound separation [26], speech enhancement [27] and source distance estimation [28]. The idea of application of EC model to sound localization was also mentioned in the work of Durlach [14], however, there has been lack of information on how this idea can be realized in practice.

3. Proposed EC-Based Sound Localization

3.1 Principle of EC-Based SSL

A null is steered to each candidate direction by using EC

operations to eliminate the signal component from that direction. Once the null is steered to the true sound source, the residual energy should be minimal. This principle was implemented in our previously-proposed SSL algorithm, namely EC-BEAM [21]. In this section, we explain it on a theoretical review point rather than focusing on technical implementation aspect.

For each interest direction θ on the horizontal plane, an equalizer is constructed so as if the source locates at θ, the signals observed at the two ear can be (or at least approximately) equalized,

$$X_L(\omega,\theta,t) - W(\omega,\theta)X_R(\omega,\theta,t) \approx 0, \tag{1}$$

where ω and t respectively denote the frequency bin index and frame index, $X_L(\omega,\theta,t)$ and $X_R(\omega,\theta,t)$ are the *short-time Fourier transforms* (STFTs) of the signals observed from the source at left and right receivers, and $W(\omega,\theta)$ is the equalizer at the direction θ. Essentially, $W(\omega,\theta)$ represents the IPD and ILD of the sound components observed from the source. In the frequency domain, the signal observed at each receiver is related to its transfer function by

$$X_i(\omega,\theta,t) = H_i(\omega,\theta)S(\omega,t), \quad i = L,R, \tag{2}$$

where $S(\omega,t)$ and $H_i(\omega,\theta)$ are respectively the sound signal emitted at the source, and the transfer function representing the propagation of sound from the source to each receiver. Therefore, the equalizer in Eq. (1) can be rewritten as

$$W(\omega,\theta) = \frac{H_L(\omega,\theta)}{H_R(\omega,\theta)}. \tag{3}$$

From Eq. (3), $W(\omega,\theta)$ is specified by only the transfer functions at θ and independent of sound signals. In the case the receivers are integrated in a dummy head, $H_i(\omega,\theta)$ becomes the *head-related* transfer function and $W(\omega,\theta)$ is equivalent to the concept of *interaural transfer function* in binaural hearing studies [15].

The equalizers are obtained by pre-training in anechoic condition where only one sound source is present at each interest direction respectively. Each equalizer is constructed independently in frequency bands, which is consistent with the modified EC model suggested by Culling and Summerfield [23]. The signal-independence property of equalizer guarantees that an equalizer trained with some sound signal is able to perform with other unknown sound signals. In training, the equalizers are calibrated using *normalized least mean square* (NLMS) method, which is given by

$$W_{t+1} = W_t + \mu\frac{X_R^*(t)}{|X_R(t)|^2}\left[X_L(t) - W_tX_R(t)\right], \tag{4}$$

where ω and θ are omitted for easy reading, the superscript * denotes the conjugate operator and μ is a scalar value specifying the step size to update the value of equalizer at each frame t until it is converged. In experiment, μ is set to 0.01.

In the stage of DOA estimation, in order to localize the sound source at an unknown direction ϕ, a null is steered to each candidate direction by using EC operations to eliminate the target signal components. After a cancellation process through all candidates, direction of sound source is specified as the direction at which the residual energy of the null is minimal, as shown in Fig. 1. That is

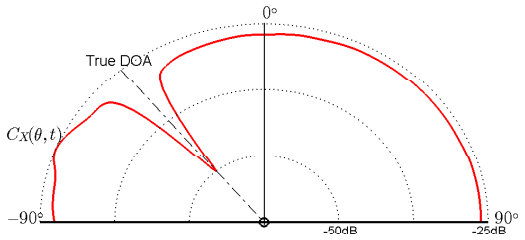

Fig. 1 Illustration of null steering in EC-BEAM. Target source locates at −40° in clean-anechoic condition.

$$\widehat{\phi}(t) = \underset{\theta}{\operatorname{argmin}}\, C_X(\theta,t), \quad \text{where}$$
$$C_X(\theta,t) = \sum_{\omega}|X_L(\omega,\phi,t) - W(\omega,\theta)X_R(\omega,\phi,t)|^2. \tag{5}$$

3.2 Robust EC-BEAM in Noisy Reverberant Conditions

In ideal condition, residual energy of the null should drop to minimum when the steering direction reaches the direction of target source. However, this may not hold when noise and reverberation are present. In practice, observed signal may consist of target sound signal and either noise or reverberation (hereafter t and ϕ are omitted for simplicity),

$$Y_i(\omega) = X_i(\omega) + R_i(\omega) + N_i(\omega), \quad i = L,R, \tag{6}$$

where $Y_i(\omega)$ is the sound signal observed at each receiver in noisy reverberant conditions, $X_i(\omega)$ is the direct-path component (target signal), $R_i(\omega)$ is the total reverberated components via indirect paths, and $N_i(\omega)$ represents the total noise in room (including its reverberation), which is assumed as uncorrelated with $X_i(\omega)$ and $R_i(\omega)$. Reverberation is built up from multiple reflections, in which each reflection can be considered as a delayed and decayed instance of the direct-path signal, that is

$$R_i(\omega) = \int_{\tau>0}^{\infty} X_i(\omega)\alpha_i(\tau)e^{-j\omega\tau}d\tau, \tag{7}$$

in which τ and $\alpha_i(\tau)$ are respectively the time delay after the direct-path and the decay coefficient due to the absorption of air, walls and other objects in room. The cancellation operation in these conditions is performed by

$$C_Y(\theta) = \sum_{\omega}|Y_L(\omega) - W(\omega,\theta)Y_R(\omega)|^2. \tag{8}$$

Since the total effect of noise and reverberation on this operation is complicated, we consider two scenarios in which noise and reverberation are treated separately.

3.2.1 Robust EC-BEAM against Noise

For simplicity, reverberation component is omitted in this

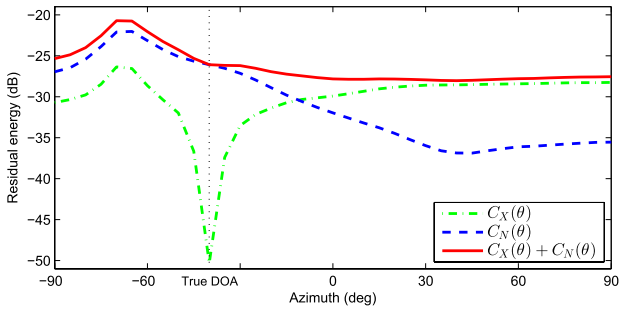

Fig. 2 Outputs of C operation with target, noise, and noisy signal. Target source locates at −40°; noise consists of diffuse noise and directional noise (at 40°) with equivalent energies; SNR of noisy signal is 0 dB.

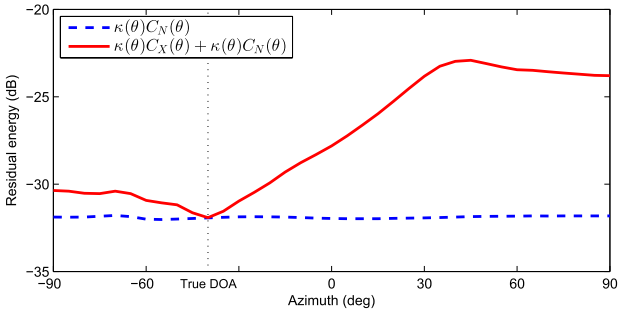

Fig. 3 Outputs of the C operation incorporated with noise compensating coefficient $\kappa(\theta)$ performed on noise and noisy signal. Configuration is same as in Fig. 2.

scenario. Observed signal in Eq. (6) is represented as

$$Y_i(\omega) = X_i(\omega) + N_i(\omega), \quad i = L, R, \tag{9}$$

in which $N_i(\omega)$ may include diffuse noise and directional noise. With the assumption of uncorrelation between target signal and noise, the C operation in Eq. (8) can be rewritten as follows (see Appendix for further explanations):

$$\begin{aligned} C_Y(\theta) &= \sum_{\omega} |X_L(\omega) - W(\omega,\theta)X_R(\omega)|^2 \\ &\quad + \sum_{\omega} |N_L(\omega) - W(\omega,\theta)N_R(\omega)|^2 \\ &\triangleq C_X(\theta) + C_N(\theta). \end{aligned} \tag{10}$$

In Eq. (10), the C operation is applied to not only target signal but also noise. Due to the compensation of $W(\omega,\theta)$, the residual noise $C_N(\theta)$ varies along the steering direction θ and affects the final output of the C process, especially when the noises at two channels are correlated (directional noise). Figure 2 demonstrates an example of this effect. Because of the considerable variation of $C_N(\theta)$, although the residual of target, $C_X(\theta)$, yields a quite good minimum to specify the DOA, the total output of C operation on both target and noise is not minimal at the direction of sound source.

In order to adapt to the effect of noise on the C operation, we additionally use a noise compensation coefficient $\kappa(\theta)$ for each steering direction θ so as the cancellation outputs of noise at all steering directions are equalized, that is

$$\kappa(\theta_m)C_N(\theta_m) = \kappa(\theta_n)C_N(\theta_n), \quad \forall m, n. \tag{11}$$

In this manner, $\kappa(\theta)$ is a scalar value characterizing for the distribution of noise energy at each direction. The coefficient $\kappa(\theta)$ specified as in Eq. (11) is independent of noise level. In implementation, $\kappa(\theta)$ is calibrated before DOA estimation by first performing the C operation with non-target signal period (where only noise is present) to obtain $C_N(\theta)$ at all directions, then by setting $\kappa(\theta)$ to 1 at 0°, i.e. $\kappa(0°) = 1$, $\kappa(\theta)$ at other directions are specified by

$$\kappa(\theta) = \frac{C_N(0°)}{C_N(\theta)}, \quad \forall\theta. \tag{12}$$

The C operation of EC-BEAM with consideration of noise, named as EC-BEAM/N, is suggested as follows:

$$\begin{aligned} C_Y^N(\theta) &= \kappa(\theta) \sum_{\omega} |Y_L(\omega) - W(\omega,\theta)Y_R(\omega)|^2 \\ &= \kappa(\theta)C_X(\theta) + \kappa(\theta)C_N(\theta). \end{aligned} \tag{13}$$

Since the equalizers satisfy Eq. (1), the output of target cancellation should drops to approximately zero when the steering direction matches the direction of target source. As a result, the residual energy of the null at the direction of target source consists of only the energy of (equalized) residual noise, i.e. $C_Y^N(\phi) = \kappa(\phi)C_N(\phi)$, and the minimum is yielded at the direction of sound source, as illustrated in Fig. 3.

It is clear that EC-BEAM/N can work well with noise as long as the coefficients $\kappa(\theta)$ are properly obtained. In fact, $\kappa(\theta)$ can be constructed for background noise even in the case it contains directional noise, provided that the noise sources are fixed and the energy of each source is relatively stable in time. Such kind of noise is popular in normal room conditions, for example the noise from fans and air-conditioners. However, this strategy is not able to deal with reverberation because its effect cannot be learned in the absence of target signal.

3.2.2 Robust EC-BEAM against Reverberation

Similarly to the first strategy, in this scenario, we omit noise component for simplicity. Sound emitted in reverberant environments arrives at receivers via multiple paths because of reflection. We conceptually divide the signal component observed from the target source into two components: early response and late response:

$$\begin{aligned} Y_i(\omega) &= X_i(\omega) + R_i^E(\omega) + R_i^L(\omega) \\ &= X_i^E(\omega) + X_i^L(\omega), \quad i = L, R, \end{aligned} \tag{14}$$

in which the early response $X_i^E(\omega)$ includes the direct-path sound component $X_i(\omega)$ and the total early reflections, $R_i^E(\omega)$, which arrive within a time delay t_0 after $X_i(\omega)$, while late response $X_i^L(\omega)$ consists of all late reflections, $R_i^L(\omega)$, arriving after t_0. When reverberation is high, $X_i^L(\omega)$ is one of the serious components effecting sound localization and perception [29]. Therefore, the strategy in this scenario mainly

aims at reducing the effect of this component on EC-BEAM. Because the component $X_i^L(\omega)$ at each receiver can be assumed as uncorrelated with $X_i^E(\omega)$ and together uncorrelated at both channels, the C operation in Eq. (8) can be rewritten as follows (see Appendix):

$$C_Y(\theta) = \sum_{\omega} \left| X_L^E(\omega) - W(\omega,\theta) X_R^E(\omega) \right|^2 + \sum_{\omega} \left[|X_L^L(\omega)|^2 + |W(\omega,\theta)|^2 |X_R^L(\omega)|^2 \right]. \quad (15)$$

Equation (15) shows that cancellation output of late response component varies along the steering directions only because of the amplitude of the equalizers, which corresponds to the ILD of target signals. Therefore, the C operation of suggested EC-BEAM against late reverberation, named as EC-BEAM/R, is executed without ILD compensation as follows:

$$\begin{aligned} C_Y^R(\theta) &= \sum_{\omega} \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} - \frac{W(\omega,\theta)}{|W(\omega,\theta)|} \frac{Y_R(\omega)}{|Y_R(\omega)|} \right|^2 \\ &= \sum_{\omega} \left| \frac{X_L^E(\omega)}{|Y_L(\omega)|} - \frac{W(\omega,\theta)}{|W(\omega,\theta)|} \frac{X_R^E(\omega)}{|Y_R(\omega)|} \right|^2 \\ &\quad + \sum_{\omega} \left[\left| \frac{X_L^L(\omega)}{Y_L(\omega)} \right|^2 + \left| \frac{X_R^L(\omega)}{Y_R(\omega)} \right|^2 \right]. \end{aligned} \quad (16)$$

As a result, the residual of late response component in Eq. (16) is independent of steering directions and does not affect the minimum of cancellation output. DOA is now specified based on the minimum of cancellation of the early response component only. Note that although EC-BEAM/R is proposed to deal with late reverberation, it is also effective with other interference having similar characteristic with this component, such as diffuse noise.

3.2.3 Robust EC-BEAM in Noisy Reverberant Environment

In order to estimate DOA in the presence of noise and reverberation simultaneously, we propose to integrate both strategies into EC-BEAM algorithm to improve its performance, namely Robust EC-BEAM. Final DOA estimate is decided based on the residual energies of EC-BEAM/N and EC-BEAM/R. As $C_Y^N(\theta)$ and $C_Y^R(\theta)$ vary over different ranges, we empirically combine them as follows:

$$C_{Robust}(\theta) = \lambda \log C_Y^N(\theta) + (1-\lambda) \log C_Y^R(\theta), \quad (17)$$

where λ is the combination coefficient specifying whether noise or reverberation is the important factor affecting estimation performance. In experiment, λ is set to 0.5, which indicates that noise and reverberation are equally treated.

4. Evaluation

4.1 Materials and Configuration

In evaluation, we simulate various reverberant conditions by using the ROOMSIM package [30]. The ROOMSIM software utilizes the HRTF measurements obtained by KEMAR dummy head [31] to generate simulated reverberant binaural room impulse responses (BRIRs) based on the image method [32]. BRIRs generated in this way are expected to represent reliable simulations since the software uses real measured HRTFs. Reverberant BRIRs are generated in a $10 \times 10 \times 3$ (m^3) room with reverberation times (T_{60}) from 0 to 0.8 s, depending on experiments. Anechoic BRIRs are selected directly from HRTF measurements without simulation. Source location varies from $-90°$ to $90°$ with the step of $5°$, at the distances from 1 to 4 m with the step of 1 m.

Speech data are selected from ATR Japanese database [33]. Six speech sentences with an average length of 10 seconds are chosen, in which three are uttered by males and the others are uttered by females. These speech sentences are convolved with the simulated BRIRs to generate directional sound signals. Simulated background noise is added into directional signals to produce noisy reverberant data. The background noise consists of diffuse noise and directional noise, in which diffuse noise is generated by first filtering sounds recorded from an air-conditioner by BRIRs at all directions then summing, while directional noise is created by filtering a sound from a fan with the BRIRs at $40°$. As these recorded noises have some different characteristics, the source of directional noise can still be perceived after mixing. The energies of two kinds of noise in the mixture are kept to be about the same. When adding the total noise to reverberant signals, the power of noise is controlled to obtained signal-to-noise energy ratio (SNR) from -5 dB to 15 dB (step of 5 dB).

The equalizer at each direction is trained using clean-anechoic signal generated from a speech sentence and used to estimate DOA of signals generated from other five sentences in all the experiment conditions. Training process is conducted using NLMS method as specified in Eq. (4). The noise compensation coefficient $\kappa(\theta)$ is calibrated as descriptions in Sect. 3.2 by using one-second period of noise. As $\kappa(\theta)$ is independent of noise level, it is calibrated only one time and applied to all SNR conditions. In test, we use a window length of 0.1 s with 50% overlapping and integrate the response energy over 0.5 s for each estimate. We evaluate the performance of SSL via the ratio of incorrect estimates to all estimates (*error rate*), where an incorrect estimate is defined as one having absolute error over $10°$.

4.2 Experiment 1: Effectiveness of Improving Strategies

In this experiment, we evaluate the effectiveness of the strategies to improve EC-BEAM proposed in Sect. 3.2. Four algorithms are examined, including the original EC-BEAM, the EC-BEAMs using individual strategies to deal with noise (EC-BEAM/N) and reverberation (EC-BEAM/R) and the EC-BEAM combining both strategies (Robust EC-BEAM). Test data are generated following the descriptions in Sect. 4.1 with a source distance of 3 m.

Figure 4 and Fig. 5 respectively demonstrate the im-

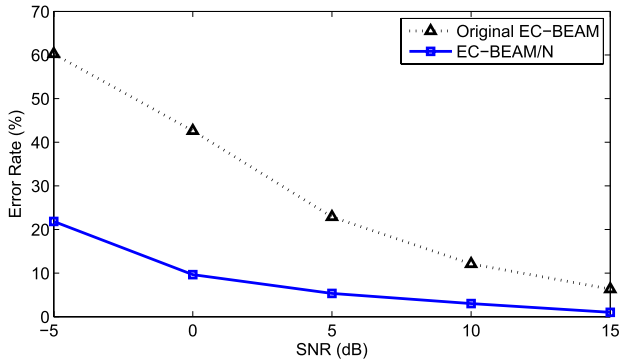

Fig. 4 Performance of original EC-BEAM and EC-BEAM/N in noisy anechoic condition. Sound source locates at the distance of 3 m.

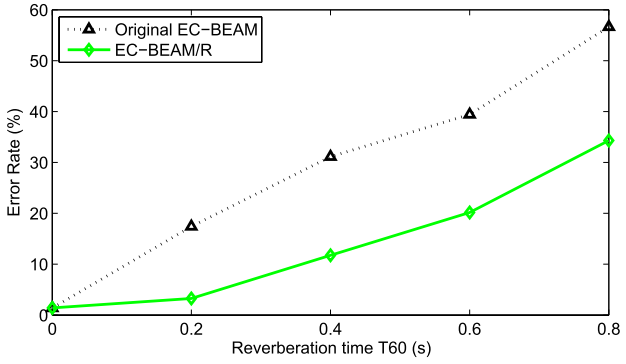

Fig. 5 Performance of original EC-BEAM and EC-BEAM/R in reverberant conditions. Source distance is at 3 m and no noise is present.

provements of EC-BEAM/N and EC-BEAM/R in comparison with the original EC-BEAM. Figure 4 shows the error rates of the original EC-BEAM and EC-BEAM/N along SNRs in anechoic condition. The original EC-BEAM can localize sound source relatively well at low noise conditions. However, its error rate rapidly increases as noise level gets higher. This is due to the effect of noise as analyzed in Sect. 3.2.1. By learning and adapting to the effect of noise, the error rate of EC-BEAM/N is dramatically reduced. Similar results are observed in Fig. 5, which represents the error rates of original EC-BEAM and EC-BEAM/R along reverberation time in the absence of noise. We can see that the original EC-BEAM also suffers from the effect of reverberation while the EC-BEAM/R is quite robust against this effect. These results indicate that each strategy works efficiently in the condition that it is designed for, i.e. noise or reverberation is present individually.

We further examine both strategies and their combination in the conditions where both noise and reverberation are concurrently present. Figure 6 shows the error rates of the four algorithms in the case $T_{60} = 0.5$ s and SNR varies from -5 dB to 15 dB. It can be observed that the performances of the original EC-BEAM, EC-BEAM/N and EC-BEAM/R are degraded in comparison with their performances in the conditions where noise or reverberation is present alone. However, EC-BEAM/N and EC-BEAM/R still work con-

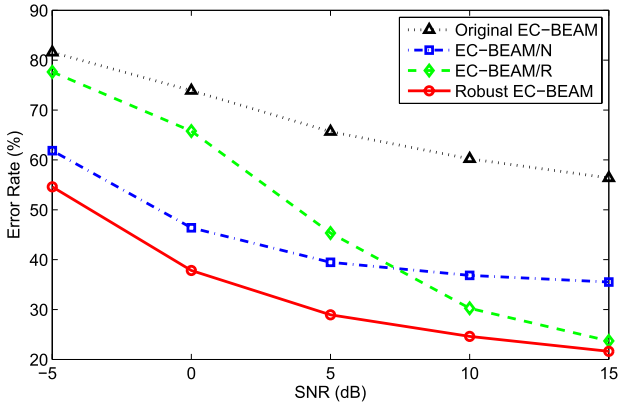

Fig. 6 Performance of the four algorithms in noisy reverberant conditions. Source distance is at 3 m and $T_{60} = 0.5$ s.

siderately better the original EC-BEAM. This indicates that the strategy to deal with an interference component does not much suffer from other interference component. EC-BEAM/N performs better than EC-BEAM/R in high noise conditions as it can well adapt to noise. When the SNR is higher than 5 dB, EC-BEAM/R outperforms EC-BEAM/N as it can account for reverberation. The Robust EC-BEAM algorithm makes fully use of the advantages of both strategies and achieves the best performance through all SNR conditions. This supports that integrating the two strategies into EC-BEAM is reasonable and the Robust EC-BEAM algorithm is able to deal with noise and reverberation simultaneously.

4.3 Experiment 2: Superiority of Robust EC-BEAM

In this section, we evaluate the proposed Robust EC-BEAM algorithm in various noisy reverberant conditions and further compare it with the standard SRP-PHAT algorithm [2] and the HRTF-based algorithm, namely Cross HRTF, introduced in [11]. Reverberation time is set to 0.5 s, while distance of sound source varies from 1 to 4 m (step of 1 m). In implementation, the SRP-PHAT method uses IPDs calculated in anechoic condition at each direction to perform beamforming. The Cross HRTF method is executed using directly the HRTFs measured in anechoic condition.

Figure 7 shows the average error rates of Cross HRTF, SRP-PHAT and Robust EC-BEAM respectively across all distances. It can be observed that the Cross HRTF method yields highest error rates among the three algorithms. The Cross HRTF has the advantage that it possesses the HRTFs, which provide the propagation information at each candidate direction. Therefore, it was reported with relatively accurate estimation in low interference (noise and reverberation) conditions [11]. However, as it does not account for interference effects, its performance is dramatically degraded in the presence of either high noise or high reverberation. Moreover, since this method is strictly dependent on HRTFs, its applicability to practice may be limited because accurately measuring these information in an arbitrary binaural

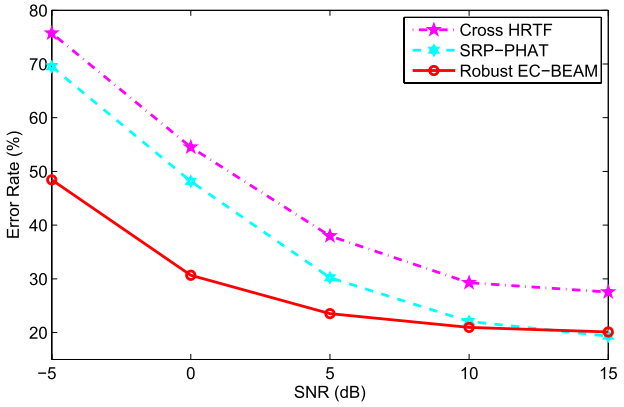

Fig. 7 Performance of Cross HRTF, SRP-PHAT and Robust EC-BEAM along SNRs. Error rate at each SNR is mean of error rates through distances from 1 m to 4 m.

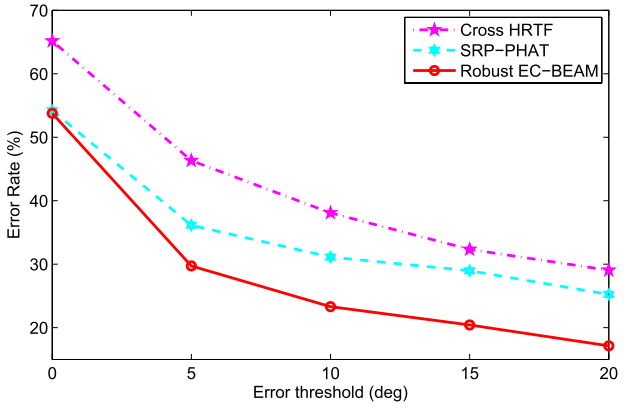

Fig. 8 Error rates of Cross HRTF, SRP-PHAT and Robust EC-BEAM along error thresholds at the fixed 5-dB SNR. Error rate at each threshold is mean of error rates through distance from 1 m to 4 m.

system is a time-consuming work.

The error rate of SRP-PHAT is consistently lower than that of Cross HRTF in both high and low noise conditions. This is partly because SRP-PHAT is quite robust against reverberation since it was shown as an approximation of maximum likelihood in low noise condition [4]. However, it still suffers from high noise condition as its error rate dramatically increases at low SNRs. This implies that SSL methods based on ITD (or IPD) only may not be fully adequate to binaural systems because the ILD, which varies largely through the azimuths due to the effect of HRTFs, is also a very important cue to specify source direction.

The Robust EC-BEAM algorithm outperforms both Cross HRTF and SRP-PHAT, especially in high noise conditions. At the SNR of −5 dB, the proposed method improves roughly 20% and 25% error rate comparing to SRP-PHAT and Cross HRTF, respectively. In relatively low noise conditions when SNR is higher 10 dB, our method performs equivalently to SRP-PHAT but still improves about 8% error rate in comparison with Cross HRTF. This is because both noise and reverberation are taken into consideration in the proposed method. The noise adaption strategy makes EC-BEAM more robust against background noise, while the strategy for reverberation significantly reduces the effect of this factor, especially the late reflection component. Besides, our method is more flexible than the Cross HRTF method since training the equalizers with observed signals should be easier than meticulously measuring HRTFs.

Figure 8 shows the error rates of the three algorithms along the error thresholds in a typical case where SNR is fixed at 5 dB. It can be observed that the error rate of Robust EC-BEAM is lower than those of Cross HRFT and SRP-PHAT at not only the defined 10°-threshold but also at all of other thresholds. This implies that our proposed method is still better than the others in the case higher accuracy is required, and the angular error of our wrong estimates (whose error is higher than 10°) is also smaller than those of the other algorithms. Figure 9 further compares the robustness of the three algorithms against reverberation. In the

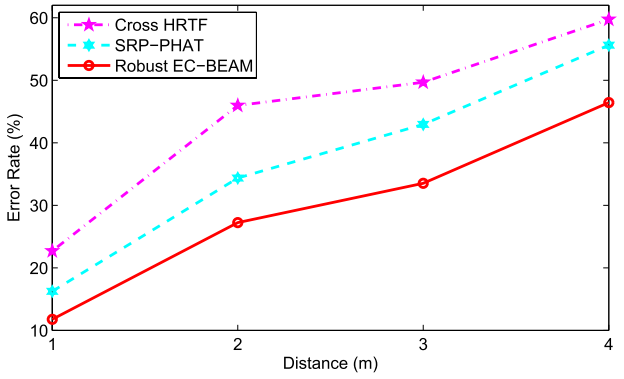

Fig. 9 Performance of Cross HRTF, SRP-PHAT and Robust EC-BEAM along distances. Error rate at each distance is mean of error rates through SNRs from −5 dB to 15 dB.

same room condition, the energy of reverberation increases relatively to that of direct component when the distance of sound source increases [34]. As a result, the effect of reverberation on signals received from longer-distance source is also higher. Observation from Fig. 9, the error rate of Cross HRTF rises quickly through distances, indicating that this method quite suffers from reverberation. Both error rates of SRP-PHAT and Robust EC-BEAM increase slower than that of Cross HRTF, in which the error rate of Robust EC-BEAM is always below that of SRP-PHAT because it is able to adapt to noise (including directional noise) in the room.

5. Discussion

The general principle of SSL is to exploit localization cues to determine direction of sound source. Therefore, methods in binaural SSL differ from each other by the way of how binaural cues are utilized. On mathematical view point, these methods have some equivalences. The Cross HRTF method [11] relies on the following principle:

$$x_L(\phi,t)\circledast h_R(\phi) = h_L(\phi)\circledast s(t)\circledast h_R(\phi) = h_L(\phi)\circledast x_R(\phi,t), \quad (18)$$

where $\circledast$ denotes convolution operator, $x_i(\phi, t)$ and $h_i(\phi)$ are respectively the observed signal and the transfer function (in the time domain) at the receiver i ($i = L, R$), provided that the source $s(t)$ locates at the direction ϕ. As a result, DOA is specified via looking for the pair of HRTFs minimizing the dissimilarity of the cross HRTFs, that is

$$\widehat{\phi}(t) = \underset{\theta}{\operatorname{argmin}}\, e(\theta, t), \quad \text{where}$$
$$e(\theta, t) = \sum_t [x_L(t) \circledast h_R(\theta) - x_R(t) \circledast h_L(\theta)]^2. \tag{19}$$

Equation (19) can be rewritten in the frequency domain as follows:

$$E(\theta) = \sum_\omega |X_L(\omega)H_R(\omega, \theta) - X_R(\omega)H_L(\omega, \theta)|^2$$
$$= \sum_\omega \left| H_R(\omega, \theta) \left[X_L(\omega) - \frac{H_L(\omega, \theta)}{H_R(\omega, \theta)} X_R(\omega) \right] \right|^2. \tag{20}$$

From Eqs. (3), (5), (20), it can be realized that Cross HRTF is a filtered version of original EC-BEAM, where $H_R(\omega, \theta)$ is the filter in this manner. Because of the equivalence of two methods, the Cross HRTF method would face similar problems with that of original EC-BEAM discussed in Sect. 3.2. This is the reason why Cross HRTF performs poorly in high noise and high reverberant conditions. Since the Cross HRTF method strictly relies on HRTFs, it is hard to understand the effect of undesired factors on this method and it has less chance to be improved. From this point of view, it would be interesting to investigate whether the proposed strategies in this paper can improve Cross HRTF in noisy reverberant environments.

In terms of sound localization using ITD, the Robust EC-BEAM against reverberation in Sect. 3.2.2 (EC-BEAM/R) has a close relation with SRP-PHAT. The generalized SRP-PHAT maximizes the sum of weighted cross correlation between each pair of N received signals, that is

$$P(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_\omega \frac{Y_i(\omega) Y_j^*(\omega)}{|Y_i(\omega) Y_j^*(\omega)|} e^{j\omega(\tau_i - \tau_j)}$$
$$= \sum_\omega \left| \sum_{i=1}^{N} \frac{Y_i(\omega)}{|Y_i(\omega)|} e^{j\omega\tau_i} \right|^2. \tag{21}$$

In the case of two-microphone, SRP-PHAT becomes

$$P(\theta) = \sum_\omega \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} e^{j\omega\tau_L} + \frac{Y_R(\omega)}{|Y_R(\omega)|} e^{j\omega\tau_R} \right|^2$$
$$= \sum_\omega \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} + \frac{Y_R(\omega)}{|Y_R(\omega)|} e^{j\omega(\tau_R - \tau_L)} \right|^2. \tag{22}$$

On the other hand, from Eq. (1), the phase component of the equalizer is the IPD of target signal. Omitting the fact that the time delay τ_i are frequency-dependent, we have

$$\frac{W(\omega)}{|W(\omega)|} = e^{j\omega(\tau_R - \tau_L)}. \tag{23}$$

The EC-BEAM/R in Eq. (16) can be rewritten as follows:

$$C_Y^R(\theta) = \sum_\omega \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} - \frac{Y_R(\omega)}{|Y_R(\omega)|} e^{j\omega(\tau_R - \tau_L)} \right|^2. \tag{24}$$

From Eq. (22) and Eq. (24), we can see that SRP-PHAT and EC-BEAM/R are quite similar. The only difference between the two methods is that one is based on the similarity of observed signals at two channels by maximizing the beamformer, the other is based on their dissimilarity by minimizing the null. These two methodologies are mentioned as the equivalent approaches in DOA estimation [15] and should provide similar results. The advantage of the proposed Robust EC-BEAM method in comparison with SRP-PHAT is the strategy accounting for the effect of noise in rooms, i.e. EC-BEAM/N. However, the this strategy is mainly effective in high noise conditions. This explains why the proposed Robust EC-BEAM algorithm localizes well at low SNRs while its performance remains as good as that of SRP-PHAT in the presence of reverberation at high SNRs.

Concerning to sound localization in a binaural system, the effects on signal observed at each receiver may include internal effect (such as HRTFs) and external effect (noise and reverberation in rooms). An efficient SSL method on such systems should be able to account for these effects. Although SRP-PHAT performs quite well in normal microphone array, this method may not be adequate to the present system as it does not account for the internal shadow effect and is low robust with noise. The methods based on HRTFs, such as cross-channel HRTFs of McDonald [11] and inverse HRTFs of Keyrouz [10], would be able to effectively exploit the internal effect. However, these methods are strictly dependent on HRTFs and do not have a mechanism to account for noise and reverberation. The method of Woodruff [9] considered these effects by building a binaural statistical model to exploit both binaural cues and monaural cues as well as to account for noise and reverberation. Nevertheless, its applicability is limited by the assumption of knowing signal-to-residual energy ratio, which is normally not available in real conditions. The algorithm proposed in this paper is able to deal with the above effects by two strategies, in which one is to learn and adapt to the effect of noise under HRTFs, the other is to account for reverberation. The adaptation strategy is similar to the mechanism that human learns the effect of reverberant room, which is mentioned as 'room learning' concept in the research of Shin-Cunningham [22]. In addition, the assumption of knowing a short period of noise in room would be reasonable since estimation of such period is possible in practice.

6. Conclusion

Sound source localization in a binaural system in practical environments is a challenging problem, since the observed signals are corrupted by either noise or reverberation while the input is limited to only two channels under HRTF effect. In the psychoacoustic research field, one of the models to explain the mechanism of human perception in such

conditions is the equalization-cancellation model. In this research, we proposed a binaural SSL method based on EC model for application in noisy reverberant conditions. Two different strategies were suggested to make the proposed method robust against noise and reverberation: one is to learn and adapt to the effect of noise under HRTFs, the other is to reduce the effect of reverberation. Experimental results showed that the proposed algorithm, which integrates both strategies, outperformed the SRP-PHAT and cross-channel HRTFs methods, and is promising for localization in adverse binaural systems in noisy reverberant environments.

Acknowledgements

The authors wish to thank Dr. Dongwen Ying in Institute of Acoustics, Chinese Academy of Sciences for his useful comments and discussions.
This study was supported by:

- The A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).
- The Strategic Information and Communications R & D Promotion Programme (SCOPE; 131205001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] L. Calmes, Biologically Inspired Binaural Sound Source Localization and Tracking for Mobile Robots, Ph.D. thesis, AWTH Aachen University, Germany, 2009.
[2] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Robust Localization in Reverberant Rooms," in Microphone Arrays, pp.157–180, Springer, 2001.
[3] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-24, no.4, pp.320–327, 1976.
[4] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in low noise, reverberative environments?," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2565–2568, 2008.
[5] R.M. Stern, G.J. Brown, and D.L. Wang, "Binaural sound localization," in Computational Auditory Scene Analysis: Principles, Algorithms and Applications, pp.147–185, Wiley, New York, 2006.
[6] S. Andersson, A. Handzel, V. Shah, and P. Krishnaprasad, "Robot phonotaxis with dynamic sound-source localization," Proc. 2004 IEEE International Conference on Robotics and Automation, pp.4833–4838, 2004.
[7] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," IEEE Trans. Audio Speech Language Process., vol.18, no.1, pp.68–77, 2010.
[8] E. Berglund, J. Sitte, and G. Wyeth, "Active audition using the parameter-less self-organising map," Autonomous Robots, vol.24, no.4, pp.401–417, 2008.
[9] J. Woodruff and D.L. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," IEEE Trans. Audio Speech Language Process., vol.20, no.5, pp.1503–1512, 2012.
[10] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3D localization based on HRTFs," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.341–344, 2006.
[11] J.A. MacDonald, "A localization algorithm based on head-related transfer functions," J. Acoust. Soc. Am., vol.123, no.6, pp.4290–4296, 2008.
[12] L.A. Jeffress, "A place theory of sound localization," J. Comparative and Physiological Psychology, vol.41, pp.35–39, 1948.
[13] N.I. Durlach, "Equalization and cancellation theory of binaural masking level differences," J. Acoust. Soc. Am., vol.35, no.8, pp.1206–1218, 1963.
[14] N.I. Durlach, "Binaural signal detection: Equalization and cancellation theory," in Foundations of Modern Auditory Theory, pp.371–462, Academic Press, New York, 1972.
[15] H.S. Colburn and A. Kulkarni, "Model of sound localization," in Sound Source Localization, pp.276–316, Springer, 2005.
[16] J.H. DiBiase, A High-Accurate, Low-Latency Technique for Talker Localization in Reverberation Environments Using Microphone Array, Ph.D. thesis, Brown University, USA, 2000.
[17] X. Lv and M. Zhang, "Sound source localization based on robot hearing and vision," Proc. International Conference on Computer Science and Information Technology, pp.942–946, 2008.
[18] D. Li and S.E. Levinson, "A linear phase unwrapping method for binaural sound source localiation on a robot," Proc. 2002 IEEE International Conference on Robotics and Automation, pp.19–23, 2002.
[19] J.C. Murray, H.R. Erwin, and S. Wermter, "Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks," Neural Netw., vol.22, no.2, pp.172–189, 2009.
[20] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am., vol.25, pp.975–979, 1953.
[21] D.T. Chau, J. Li, and M. Akagi, "A DOA estimation algorithm based on equalization cancellation theory," Interspeech, pp.2770–2773, 2010.
[22] B.G. Shinn-Cunningham, "Learning reverberation: Considerations for spatial auditory displays," International Conference on Auditory Displays, pp.126–134, 2000.
[23] J.F. Culling and Q. Summerfield, "Perceptual segregation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am., vol.98, no.2, pp.785–797, 1995.
[24] J.F. Culling, M. Hawley, and R. Litovsky, "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," J. Acoust. Soc. Am., vol.116, no.2, pp.1057–1065, 2004.
[25] R. Wan, N.I. Durlach, and H.S. Colburn, "Application of extended equalization-cancellation model to speech intelligibility with spatial distributed maskers," J. Acoust. Soc. Am., vol.128, no.6, pp.3678–3690, 2010.
[26] N. Roman, S. Srinivasan, and D.L. Wang, "Binaural segregation in multisource reverberant environments," J. Acoust. Soc. Am., vol.120, no.6, pp.4040–4050, 2006.
[27] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," Speech Commun., vol.53, no.5, pp.677–689, 2011.
[28] Y.C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," IEEE Trans. Audio Speech Language Process., vol.18, no.7, pp.1793–1805, 2010.
[29] N.I. Durlach, "Source separation, localization, and comperhension in human, machines, and human-machine systems," in Speech Separation by Humans and Machines, ch. 15, pp.221–243, Kluwer Academic, 2005.
[30] D.R. Campbell, The ROOMSIM User Guide (v3.4), 2004. Available at: http://media.paisley.ac.uk/˜campbell/Roomsim/ Accessed in December, 2013.
[31] W.G. Gardner and K.D. Martin, "HRTF measurements of a KEMAR," J. Acoust. Soc. Am., vol.97, no.6, pp.3907–3908, 1995.
[32] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol.65, no.4,

pp.943–950, 1979.
[33] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as atool of speech recognition and synthesis," Speech Commun., vol.9, no.4, pp.357–363, 1990.
[34] A.W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," Nature, vol.397, no.6719, pp.517–520, 1999.

Appendix

Given two signals **X** and **Y** in the frequency domain: $\mathbf{X} = [X(\omega_1),\ X(\omega_2), \ldots, X(\omega_n)]$, $\mathbf{Y} = [Y(\omega_1),\ Y(\omega_2), \ldots, Y(\omega_n)]$. If **X** and **Y** are uncorrelated, then

$$\sum_{\omega} |X(\omega) + Y(\omega)|^2 = \sum_{\omega} \left[|X(\omega)|^2 + |Y(\omega)|^2 \right]. \qquad (\mathrm{A}\cdot 1)$$

Proof: Since (**X**,**Y**) is uncorrelated, (**X**,**Y***) and (**X***,**Y**) are also uncorrelated, where the superscript * denotes the conjugate operator. That means

$$\sum_{\omega} X(\omega)Y(\omega) = \sum_{\omega} X(\omega)Y^*(\omega) = \sum_{\omega} X^*(\omega)Y(\omega) = 0.$$

As a result,

$$\begin{aligned}
\sum_{\omega} |X(\omega)+Y(\omega)|^2 &= \sum_{\omega} [X(\omega)+Y(\omega)][X(\omega)+Y(\omega)]^* \\
&= \sum_{\omega} [X(\omega)+Y(\omega)][X^*(\omega)+Y^*(\omega)] \\
&= \sum_{\omega} \left[X(\omega)X^*(\omega)+Y(\omega)Y^*(\omega) \right] \\
&= \sum_{\omega} \left[|X(\omega)|^2+|Y(\omega)|^2 \right].
\end{aligned}$$

Thanh-Duc Chau received the B.S. degree in Computer Science from Ho Chi Minh University of Science, Vietnam National University (HCMUS-VNU) in 2007 and the M.S. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2010. He is currently a Ph.D. candidate in JAIST. His research interests include binaural signal processing, speech enhancement and sound source localization. He received the Student Paper Award from NCSP11 in 2011.

Junfeng Li received the B.E. degree from Zhengzhou University and the M.S. degree from Xidian University, China, both in Computer Sciences, in 2000 and 2003, respectively. He received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in March 2006. From April 2006, he was a post-doctoral research fellow at Research Institute of Electrical Communication (RIEC), Tohoku University. From April 2007 to July 2010, he was an Assistant Professor in the Graduate School of Information Science, JAIST. Since August 2010, he has been a Professor in Institute of Acoustics, Chinese Academy of Sciences. His research interests include speech information processing, intelligibility hearing aids, spatial hearing and acoustic signal processing. Dr. Li received the Best Student Award in Engineering Acoustics First Prize from the Acoustic Society of America in 2006, and the Best Paper Award from JCA2007 in 2007.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) and is now a full professor. His research interests include speech perception, the modeling of speech perception mechanisms in human beings, and the signal processing of speech. During 1998, he was associated with the Research Laboratories of Electronics at MIT as a visiting researcher, and in 1993 he studied at the Institute of Phonetics Science at the University of Amsterdam. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the International Speech Communication Association (ISCA). Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, the Best Paper Award from the Research Institute of Signal Processing in 2009, and the Sato Prize for Outstanding Papers from the ASJ in 1998, 2005, 2010 and 2011.